

Effects of Timing on Users' Agency during Mixed-Initiative Interaction

Guo Yu
Computer Laboratory
University of Cambridge
15 JJ Thomson Avenue
Cambridge, CB3 0FD, UK
Guo.Yu@cl.cam.ac.uk

Alan Blackwell
Computer Laboratory
University of Cambridge
15 JJ Thomson Avenue
Cambridge, CB3 0FD, UK
Alan.Blackwell@cl.cam.ac.uk

We explore the role of timing in situations where a human user and semi-autonomous software can each initiate actions, building on cognitive theories of rhythmic expectation and mutual temporal adaptation during conversation. Two controlled experiments demonstrate that adjustments to the rhythm of back-and-forth interaction have significant effects on perceived agency, task performance and stress. Conclusions include design guidance that establishing a predictable rhythm of interaction is likely to be beneficial for mixed initiative systems.

Keywords: Mixed-Initiative Interaction, Sense of agency, Rhythm, Expectation

1. INTRODUCTION

Intelligent user interfaces increasingly complete our actions, or even act on our behalf. They range from Programming-by-Example systems that observe our actions in order to automate them [1], predictive text that anticipates our next word [2], search boxes that guess the question we will ask, and semi-autonomous vehicle navigation systems that tell us when to turn the steering wheel, or even turn it by themselves [3]. Such systems acquire 'mixed initiative' characteristics - sometimes the user takes the initiative, and sometimes the system does [4]. Nevertheless, in order to be usable, such systems must allow the user to maintain a 'locus of control' [5] - Shneiderman's term for relating the system behaviour to the user's intentions. Importantly, such control is reflected in a 'sense of agency' [6]. Sense of agency reflects the extent to which a person feels themselves to be in control and have influence over his or her own actions, and is fundamental to mental health and social wellbeing.

In this research, we are concerned with design factors that influence the sense of agency in mixed initiative systems, and ways in which agency can be measured as an aspect of user experience. We are particularly interested in the ways that timing of mixed initiative interaction might emulate interaction between two humans. Inappropriate timing of human interactions is reflected in expressions such as 'he jumped down my throat' to describe a person

who takes the initiative in conversation faster than appropriate. Until now, studies of timing in HCI have been influenced by real time systems engineering. According to that perspective, we want user interfaces to respond as fast as possible, but have not considered the possible dangers when they respond *too* fast. Our research question is therefore to investigate what timing characteristics would be most appropriate for mixed initiative interaction.

2. RELATED WORKS

2.1. Agency in human-computer interaction

The study of locus of control (as in [5]) builds on earlier philosophical and psychological theories [7; 8], in which perceived control is described as 'experience of agency' [9; 10]. A person will have a sense of agency when they consider themselves to have the ownership of, and be responsible for, the consequences that their actions have in the external world [6; 11]. Cognitive neuroscientists take one of two stances in explaining how the sense of agency arises. The first is called the *Comparator Model*, which maintains that people will experience a sense of agency when the actual sensory consequences match with the prediction made by the motor system. The alternative suggests that sense of agency arises from *retrospective inference*, based on the *Apparent Mental Causation Model*. This model maintains that people infer a causal link if three criteria are met:

a) the action occurred prior to the outcome; b) the outcome was consistent with expectation; and c) this action was the only plausible cause of the outcome. Previous research in HCI has studied the concept of agency from several perspectives [6]. The first can be summarised as *how to take actions*, and focuses on how different input modalities (e.g. speech, gesture, or skin input) affect users' sense of control [11; 14; 15; 16]. A second focuses on *how to present consequences*, comparing alternative output modalities (e.g. visual, audio or haptic feedback) [17; 18; 19]. Our own research explores *how actions and consequences are aligned*, because in mixed-initiative interaction, the back-and-forth flow involves constant transitions between 'user initiates, computer responds' and 'computer initiates, user responds'. We aim to find design approaches that allow users to preserve a sense of agency during these exchanges of initiative. Although there are many parameters that link actions to consequences, we are particularly interested in the timing of the interaction. Timing is a fundamental property of all interaction, both in social interaction between humans, and in their natural interaction with the physical world.

2.2. Temporal expectation

Research into the temporal experience of causality suggests that the perceived timing of actions and their consequences is adjusted to fit prior expectations [20]. Expectation based on past experience operates as a top-down process [21] that guides experiences of the self and the external world, shaping information processing as well as interpretation of other's behaviours during interaction. Furthermore, when expectation confirms a prior causal belief (that 'I' will be responsible for a consequence), it intensifies the sense of agency [22; 21].

These processes allow more efficient information processing, through encoding temporal patterns of events. Temporal expectations both enhance signal detection and facilitate pattern recognition. They result in reduced neural response to expected stimuli, but increased excitation when a signal does not appear as expected [23]. EEG analysis suggests that expectation bias enhances efficiency by constraining the interpretations of inputs to a more limited population [24; 25; 26; 27].

The degree to which expectations modulate perception of behaviour can be explained by the Expectancy Violation Theory (EVT) [28; 29; 30]. EVT suggests that a person assesses a behaviour depending on what they expect it to be. When it violates their expectation, more intensive cognitive processing is triggered to make a deeper assessment of the behaviour as well as its meaning and function. EVT offers an information-processing interpretation of the Golden Rule that users *'don't*

want surprises or changes in familiar behaviour, and they are annoyed by ... inability to produce their desired result' [5]. According to EVT, any violation of expectation, whether a positive surprise or negative frustration, might diminish users' experience of agency. In mixed-initiative interactions where it is typically assumed that automated intervention will be beneficial, it seems important to explore further the effect of temporal expectation on users' sense of agency.

2.3. Entrainment during interaction

Patterns of temporal expectation are widely studied in music neuroscience, as well as conversation and language studies, as the phenomenon of *rhythm*: 'systematic patterning of sound in terms of timing, accent, and grouping' [31]. Rhythm is distinct from periodicity: while periodicity requires repetitive patterns, rhythm can refer to any predictable and systematic patterning. Rhythm can also refer to temporal patterns in other forms of signal beyond sound, including neural activity, motions, or visual perceptions.

Rhythm has been intensively studied as a static property, for example as a classifier of musical forms and language groups, or for biometric authentication. Recent studies have started to explore functional aspects, such as its emotional effects in music, persuasive effects in speech and entrainment effects in interpersonal behavioural coordination. Entrainment refers to a process in which two or more rhythmic processes adapt to each other, eventually acting in relatively stable synchrony [32], as when two or more pendulums or other oscillators 'lock up' to each other with the same period, either in exact alignment or alternation (0 or 180 degree phase).

During interpersonal interaction, entrainment can establish mutual agreement in cognitive processes involving perception, synchronisation and adjustment [33]. This can enhance intersubjectivity - 'the sharing of subjective states by two or more individuals' [34], enhancing trust and empathy as well as pro-sociality [32; 35]. Our hypothesis is that similar phenomena can be applied to the design of mixed-initiative interaction, such that users' sense of agency can be enhanced through behavioural entrainment of user and system.

3. RESEARCH FRAMEWORK

3.1. Perceived agency from predictable rhythm

If sense of agency results from retrospectively inferred patterns, then more predictable patterns will facilitate perception of control by supporting temporal expectation with minimal cognitive resources [36]. We can assess sense of agency in two ways: firstly

by simply collecting subjective ratings of perceived control, and secondly by measuring distorted time perception that results from 'intentional binding' using the Libet clock paradigm. The Libet clock implicitly measures sense of agency [37; 38] based on the research finding that people perceive an involuntary action as happening earlier than it actually did (with intentional actions perceived as later), while an unintended outcome is perceived as occurring later than the outcome of an intentional action.

Using these two measures, we hypothesise that:

H1.1: Predictable rhythm in mixed-initiative interaction will preserve users' sense of agency.

H1.2: Irregular time intervals in mixed-initiative interaction will impair users' sense of agency.

3.2. Perceived rhythmic entrainment

Because a more rhythmic pattern is more predictable, adaptation during entrainment should require less cognitive resource. Previous research suggests that entrainment can facilitate interpersonal communication by enhancing mutual awareness (i.e. a sense of '*being together*') [39], and we expect to observe this in mixed-initiative interaction. Research in mutual adaptive tapping uses auto-correlation and cross-correlation coefficients [40] to study entrainment effects. We use joint lag autocorrelation to describe the similarity between observations given a certain time lag between them. It ranges between -1 and 1, with a positive value suggesting greater tendency for temporal assimilation, whereas a negative value suggests tendency to compensation [40]. Cross-correlation measures the similarity of two interacting series as a function of the displacement of one relative to the other, with larger values indicating stronger similarity. We used windowed cross-correlation [41], with a window width corresponding to one round of the mixed-initiative interaction task.

We hypothesise that:

H2.1: Predictable rhythm in mixed-initiative interaction is more likely to induce users' entrainment behaviours.

H2.2: Irregular time intervals in mixed-initiative interaction is less likely to induce users' entrainment behaviours.

3.3. Stress and relaxation

Studies in social psychology have shown that rhythmic entrainment provides a basis for mutual trust and predictability, resulting in a sense of relaxation [42]. In mixed-initiative interaction, this may result in reduced stress and mental effort. We measured mental demand, physical demand, amount of effort devoted using the NASA Task Load Index (TLX) ratings systems [43].

We hypothesise that:

H3.1: Predictable rhythm in mixed-initiative interaction can give users a sense of relaxation.

H3.2: Irregular time intervals in mixed-initiative interaction can give users a sense of stress.

3.4. Hypotheses on task performance

Compared with random stimuli that occur at irregular times, random stimuli that occur within a rhythmic frame would be easier for users to predict and respond to. This should allow users to devote more cognitive resources to complicated tasks and stimuli. We recorded accuracy of all task responses. We also asked participants to rate how confident they were, and how successful they perceived their performance to be.

We hypothesise that:

H4.1: Predictable rhythm in mixed-initiative interaction can help users achieve better task performance and feel more confident in their own performance.

H4.2: Irregular time intervals in mixed-initiative interaction can impair users' task performance and the confidence in their own performance.

4. EXPERIMENT 1

In order to study timing effects of mixed initiative interaction in a highly controlled way, we adapted a simple type of stimulus-response experiment, in which sequences of user-initiated actions are conventionally followed by prompts initiated by the system. We modified this conventional controlled experiment by adjusting the rhythmic aspects of the system-initiated actions.

4.1. Tasks and Participants

The first experiment aimed to study how timing patterns in visual stimuli affect users' performance and sense of control. In order to mitigate bias caused by experimental demand, we told participants that this experiment would study '*how people follow various sequences of events on a screen*', not mentioning timing or rhythm. Participants were asked to do 5 types of task, each of which required multiple mouse clicks: first on an initial prompt, then on randomised shapes appearing at a sequence of target locations on the screen. After each task, they had to recall the shape that had appeared at each location by selecting it from among alternatives. Participants practiced each task for 3 rounds, then completed the main experiment in which each type of task was repeated for 30 rounds. They reported subjective ratings on sense of control and stress after completing each type of task. We recruited 22 participants, who participated in both experiments. A small gift was given in appreciation of their time. The experiment was reviewed by the ethics committee of the Cambridge University Computer Laboratory.

4.2. Independent variable and manipulation

This experiment had one independent variable, as shown in Table 1: rhythmic intervals vs. arrhythmic intervals. There were three sub-conditions under rhythmic intervals, each of which used a different method of setting the rhythm.

Independent Variable	Description	Abbreviation
Arrhythmic	System sets all irregular intervals (Computer takes the initiative A rrhythmically)	CA
Rhythmic	System sets all rhythmic intervals (Computer takes the initiative R hythmically)	CR
	User sets interval rhythm, to which the system would then align (User takes the initiative, C omputer aligns)	UC
	User sets all interval rhythm (User takes the initiative R hythmically)	UR

Table 1: Independent variable and its settings

Figure 1 illustrates the design of each type of task. The experiment always started with a preparation task (Task 0). Participants clicked 4 target crosses appearing in order at 4 locations on the screen. They were asked to click at a rate they found comfortable for 30 rounds. All between-click intervals were recorded, with the average used later to set the rhythm for Task 2. In Tasks 1 and 2, the screen first displayed 4 crosses in sequence at 4 locations on the screen, then 4 simple shapes (randomly selected from triangle, square, pentagon and circle) at the same 4 locations. Participants then had to recall which shape had been displayed at each location. In the CA condition (Task 1), the time interval between each stimulus presentation was randomised. In the CR condition (Task 2) the intervals were fixed at the average value observed in Task 0.

In the UC condition (Task 3), participants clicked on the 4 target crosses, then waited and observed the display of 4 randomised shapes (without clicking). They were then asked to recall the shapes again. The time intervals between presentation of the shapes was exactly the intervals of users' clicking on the crosses. In the UR condition (Task 4) participants clicked on 4 target crosses at the same locations, then on 4 randomised shapes, all at their own preferred rhythm. Then they needed to recall the shapes as before. The sequence of Tasks 1, 2, 3 and 4 was randomised for each participant.

4.3. Dependent variables and measurements

The timing of all interaction events was recorded as timestamps of stimulus presentations and participant mouse clicks. These were used to calculate inter-stimulus and inter-click intervals. As shown in Figure 2, there are 12 intervals in each round, falling into 3

stages: The first four are intervals before each pre-target presentation (denoted as $I(r_i, p_1)$, $I(r_i, p_2)$, $I(r_i, p_3)$, $I(r_i, p_4)$). The next four are intervals before a target presentation (denoted as $I(r_i, t_1)$, $I(r_i, t_2)$, $I(r_i, t_3)$, $I(r_i, t_4)$). The final four are intervals before an answer (denoted as $I(r_i, a_1)$, $I(r_i, a_2)$, $I(r_i, a_3)$, $I(r_i, a_4)$). We calculated three dependent variables to describe changes in rhythm over time: the auto-correlation of participants' *Answer* intervals during the answer stage of two successive rounds; the cross-correlation between *Pre-Target* intervals and *Answer* intervals within one round; and the cross-correlation between *Target* intervals and *Answer* intervals within one round.

We recorded participants' choices of shape and location during the recall stage, and calculated the dependent variable *Accuracy* as the number of correct answers in each round.

After each task, subjective measures were captured by presenting participants with two sets of sliders (initialised to the mid position), having paired opposite statements at each end. As described in section 3.3, we adopted the NASA-TLX scale to assess mental demand, physical demand, temporal demand, performance, effort and frustration. We also asked participants to rate the following 5 items:

- a) 'The software adapted to me' vs. 'I adapted to the software'
- b) 'I was controlling the pace' vs. 'The software was controlling the pace'
- c) 'The software intended to help me' vs. 'The software intended to challenge me'
- d) 'I felt relaxed during this task' vs. 'I felt stressed during this task'
- e) 'I felt confident in my answers' vs. 'I felt unconfident in my answers'

4.4. Results

4.4.1. Subjective report

To test the effectiveness of independent variable manipulation and hypotheses **H1.1**, **H1.2**, **H3.1**, **H3.2**, **H4.1** and **H4.2**, we analysed participants' subjective ratings. The data did not pass the Shapiro-Wilk Normality Test, therefore we used the non-parametric Friedman Test to analyse the effect of rhythm across 4 conditions and the Wilcoxon Signed Ranks Test for pairwise comparisons.

Participant reports on sense of control demonstrates that the manipulation of our independent variable has been effective as anticipated in **H1.1** and **H1.2**. As shown in Figure 3, a stronger sense of control of the interaction pace appeared in UR than in the other conditions (CA: $Z=-4.109$, $p<0.001$; CR: $Z=-4.107$, $p<0.001$; UC: $Z=-3.528$, $p<0.001$). The UC condition was also rated higher than CA ($Z=-3.059$, $p=0.002$) and CR ($Z=-2.433$, $p=0.015$),

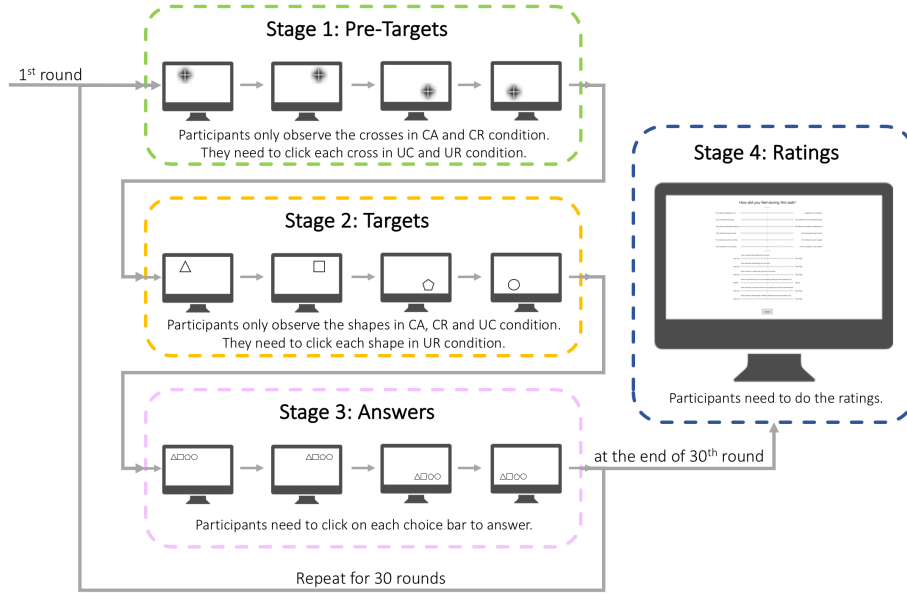


Figure 1: Illustration of Tasks in Experiment 1

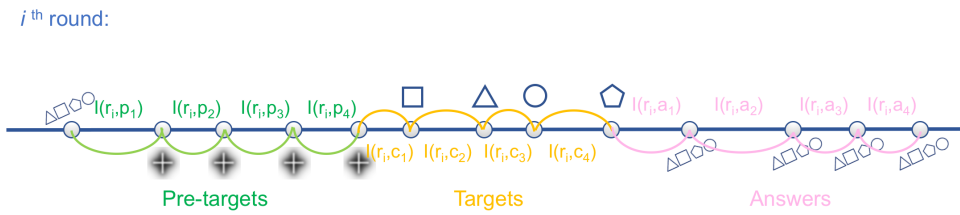


Figure 2: Stages and Intervals in Experiment 1

CR rated higher than CA ($Z=-2.017$, $p=0.044$).

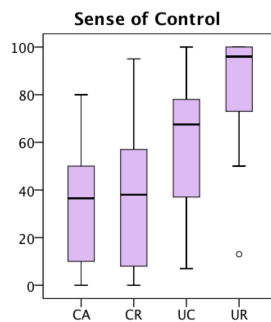


Figure 3: Ratings on 'Sense of Control' in Experiment 1

Among the 6 items in the TLX scale, significant effect was found on the ratings of physical demand ($\chi^2=12.277$, $p=0.006$), performance ($\chi^2=13.206$, $p=0.004$) and effort ($\chi^2=9.332$, $p=0.025$), see Figure 4. Participants found the physical demand in the UR condition stronger than other three (CA: $Z=-2.664$, $p=0.008$;

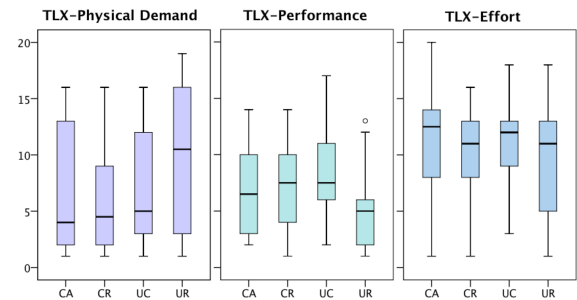


Figure 4: Ratings on TLX 'Physical Demand', 'Performance' and 'Effort' in Experiment 1

CR: $Z=-3.202$, $p=0.001$; UC: $Z=-2.401$, $p=0.016$), and rated UC as more physically demanding than UR ($Z=-2.045$, $p=0.041$). However, note that the UR and UC tasks required more clicking. The scale for performance rating was marked 'perfect' at its left end and 'failure' at the other, therefore the more successful participants consider themselves, the lower the ratings would be. Results showed that participants rated their performance

in the UR condition better than in others (CA: $Z=-1.950$, $p=0.051$; CR: $Z=-2.025$, $p=0.043$; UC: $Z=-2.954$, $p=0.003$), and this has partially supported **H4.1** and **H4.2**.

H3.1 and **H3.2** are both supported by participants' rating on effort: they considered they had devoted more effort in CA than in CR ($Z=-2.199$, $p=0.028$) and UC ($Z=-2.229$, $p=0.026$), and less effort in UR than in CR ($Z=-2.075$, $p=0.038$) or UC ($Z=-2.103$, $p=0.035$).

4.4.2. Accuracy

Hypothesis **H4.1** and **H4.2** were tested using the non-parametric Friedman Test because the numbers of correct answers were not normally distributed. Significant effects were found across the 4 conditions ($\chi^2=8.497$, $p=0.037$), see Figure 5. The accuracy in the UR condition was significantly better than in CA ($Z=-1.976$, $p=0.048$) and UC ($Z=-2.446$, $p=0.014$), and CR was marginally better than UC ($Z=-1.936$, $p=0.053$). This result has supported **H4.1** and **H4.2**.

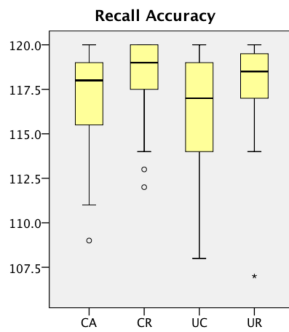


Figure 5: Participants' Recall Accuracy in Experiment 1

4.4.3. Cross-correlation and auto-correlation

In order to test Hypothesis **H2.1** and **H2.2**, we compared cross-correlation and auto-correlation coefficients using repeated measures ANOVA (having passed the Shapiro-Wilk Normality Test and the Mauchly's Sphericity Test). Because the *Pre-Targets* and *Targets* intervals within any round of the CR condition were identical, their standard deviation was always 0, and cross-correlation not relevant. In the CR condition we therefore analysed within-round cross-correlation only for the *Pre-Targets* and *Answers* intervals, see Figure 6. Paired-samples t-test revealed that the cross-correlation between *Pre-Targets* intervals and *Answers* intervals was significantly larger in the UC condition than CA ($t=7.292$, $p<0.001$) and UR ($t=4.661$, $p<0.001$), and this correlation in UR was significantly larger than CA ($t=-3.402$, $p=0.003$). A significant difference in cross-correlation of *Targets* intervals and *Answers* intervals was also found between these three conditions. Again we found the

cross-correlation in UC was significantly larger than CA ($t=8.380$, $p<0.001$) and UR ($t=5.653$, $p<0.001$), and UR cross-correlation was marginally larger than CA ($t=-1.810$, $p=0.085$). Since higher cross-correlation suggests stronger entrainment tendency, the results support **H2.1** and **H2.2**, i.e. participants entrained their *Answers* intervals with regular system intervals, but did not when system intervals were irregular.

Further analysis of auto-correlation provided strengthened support for **H2.1** and **H2.2**. The difference in auto-correlation between rhythmic and arrhythmic interaction was significant ($F=18.702$, $p<0.001$), see Figure 7. Paired-samples t-test showed that the auto-correlation of participants' free pace clicking intervals was significantly lower than the auto-correlation of participants' *Answers* intervals in each condition (CA: $t=6.212$, $p<0.001$; CR: $t=6.412$, $p<0.001$; UC: $t=4.674$, $p<0.001$; UR: $t=2.548$, $p=0.019$), and the auto-correlation in the UR condition was also significantly lower than the other conditions (CA: $t=4.950$, $p<0.001$; CR: $t=4.194$, $p<0.001$; UC: $t=3.342$, $p=0.003$). In other words, participants exhibited as much self assimilation in the CA condition as they did in the CR and UC condition, which demonstrates their struggles not to entrain with irregular system intervals.

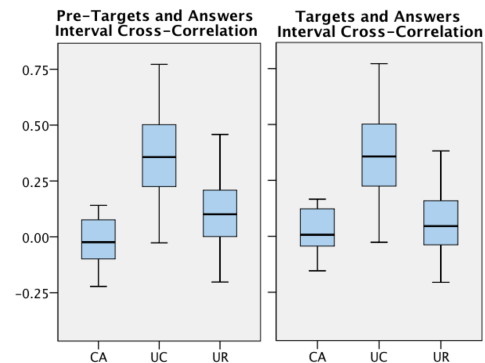


Figure 6: Average Cross-correlation (within one round) in Experiment 1

4.4.4. Discussion

In Experiment 1, when participants were manually setting the rhythm in Task 4 (UR), participants showed higher sense of control, had higher confidence in their own performance and actually did achieve higher accuracy. Despite the fact that UR was the most physically demanding task, participants still thought they had devoted the least effort in it. The implication is that during mixed-initiative interaction, greater reliance on manual control at a relatively micro level would not necessarily increase user stress, because they may enjoy being able to track their actions and outcomes. Interestingly, when participants had full control of

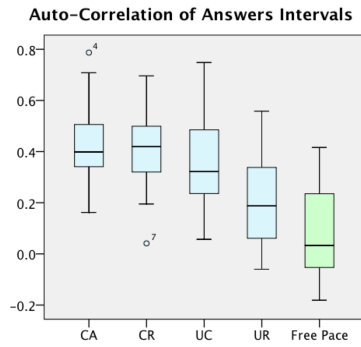


Figure 7: Average Auto-correlation of Answers Intervals (between successive rounds) in Experiment 1

pace during Task 0 (free pacing) and Task 4 (UR), they let the rhythm become looser over time, as seen from low auto-correlation of their own clicking sequence. However, when the system started to take more initiative (in Task 1, 2 and 3), it seemed that participants started to regulate their behaviour rhythm. Our interpretation of this phenomenon is that maintaining temporal regularity might be a strategy to assert control, even if just the perception of control.

It is not surprising that participants experienced the least sense of control, most effort and worst accuracy when the system set an arrhythmic pace in Task 1 (CA), as hypothesised in **H1.2**, **H2.2**, **H3.2** and **H4.2**. However, participants seemed to exhibit a high level of self assimilation, as if fighting against this unpredictability by asserting their own rhythm. This phenomenon can be seen from the analysis of auto-correlation and cross-correlation: while the tendency to self assimilation in the CA condition was as high as that of CR, participants did not entrain with arrhythmic *Pre-Targets* and *Targets* intervals in the way they did with rhythmic ones. Considering their loose pace in Task 0 and Task 4, maintaining such level of regularity may have contributed to their perceived effort.

When the system presented stimuli rhythmically in Task 2 (CR), though participants were not in control of the rhythm, their task performance was almost as good as that of Task 4 (UR). The perceived effort was also lower than when the stimuli were arrhythmic (CA) or when they had only half of the control (UC). They also showed a tendency to entrain with the rhythmic intervals, because the within round cross-correlation coefficients were the highest in the CR condition. This supports our **H2.1**, **H3.1** and **H4.1**, and also previous findings that entraining with a rhythmic external process is energy-efficient and beneficial. The design implication is that where possible, timing of system actions and responses (on a micro level) should happen regularly in time.

5. EXPERIMENT 2

The results of the first experiment support our hypotheses that predictable rhythm can preserve users' sense of agency, facilitate entrainment, reduce stress and enhance task performance. To further explore how rhythmic aspects of system-initiated actions would influence users' timing perception and sense of agency, we designed an experiment using the intentional binding paradigm. Once again, we manipulated rhythmic aspects of the interaction between the user and the system.

5.1. Tasks and Participants

Experiment 2 used the same structure as Experiment 1, but using auditory rather than visual stimuli. Participants were told that the purpose of this experiment was to explore '*how people follow various sequences of sounds from a computer*'. As before, there were 5 types of task, each of which required participants to listen to randomised number of beeps while observing a standard Libet clock [37]. They reported the position of the clock hand at the last beep by typing numbers into a text box. Figure 8 illustrates the task procedure. Participants practiced each task for 3 rounds, then 30 rounds of each task in the main experiment. They provided subjective ratings after each block as before.

5.2. Independent variable and manipulation

In Task 0, participants chose a beeping rhythm that they felt comfortable with, adjusted by dragging a slider. The system enabled a confirm button after a selected rhythm had repeated 16 times. This was used as the rhythm in Task 2. In Task 1 and Task 2, participants listened to a series of beeps while observing the Libet clock. The number of beeps could be randomly 7, 8, 9 or 10. In the CA condition (Task 1) intervals were completely irregular. In the CR condition (Task 2) all intervals were fixed as determined in Task 0. In the UC condition (Task 3), participants clicked a button to make the computer beep for 4 times, after which the computer system continued to beep for another 3, 4, 5 or 6 times (randomised). In the UR condition (Task 4), participants repeatedly clicked a button to make the computer beep, continuing until the button disappeared after either 7, 8, 9 or 10 clicks. For each round, participants reported the position of the clock hand at the last beep of that round. The sequence of Tasks 1, 2, 3 and 4 was randomised for each participant.

5.3. Dependent variables and measurements

The dependent variable in Experiment 2 was the standard measure of outcome binding used in the Libet clock paradigm, which is calculated by

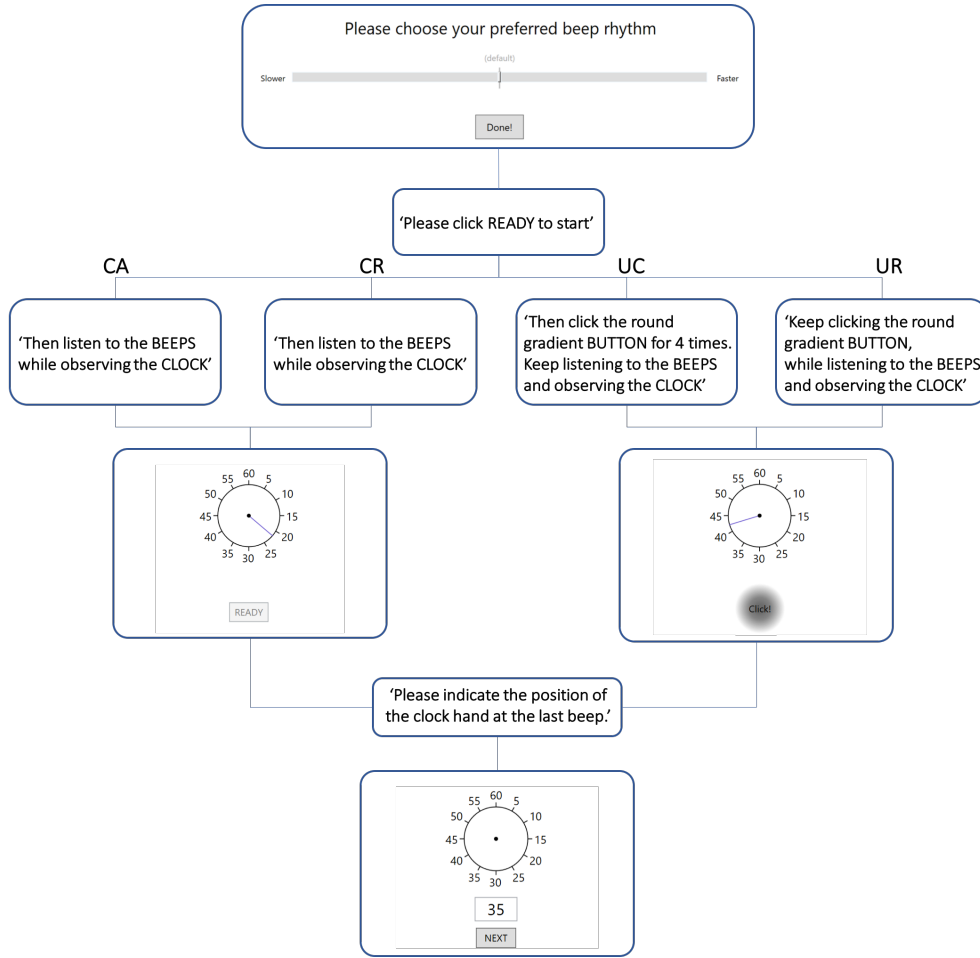


Figure 8: Illustration of Tasks in Experiment 2

subtracting the average value of participants' *active error* from the average value of their *baseline error* [6]. Baseline error is the difference between the actual time and participants' perceived time for a random beep generated by the system. Active error is the difference between the actual time and participants' perceived time of the last beep in each round. All components were measured in milliseconds. In the Libet clock paradigm, a more negative value of outcome binding effect indicates lower sense of agency. Subjective report variables were collected in the same way as for Experiment 1.

5.4. Results

5.4.1. Subjective report

Subjective ratings were not normally distributed, so we used the non-parametric Friedman Test to test **H1.1**, **H1.2**, **H3.1**, **H3.2**, **H4.1** and **H4.2**. Participants reported that the system was adapting to them more in the UC condition (CA: $Z=-3.129$, $p=0.002$; CR: $Z=-3.529$, $p<0.001$; UR: $Z=-2.334$, $p=0.020$). They also rated UR more adaptive than CA ($Z=-2.576$, $p=0.010$) and CR ($Z=-2.096$, $p=0.036$). Participants perceived

themselves to control the pace more in the UR condition than in UC ($Z=-3.665$, $p<0.001$), while UC provided more control than CA ($Z=-3.458$, $p=0.001$) and CR ($Z=-4.108$, $p<0.001$). This can be seen in Figure 9, which demonstrates that the manipulation of the independent variable in Experiment 2 is also effective, and partially supports **H1.1** and **H1.2**. Participants indicated that the system challenged them more in the CA condition (CR: $Z=-3.527$, $p<0.001$; UC: $Z=-3.463$, $p=0.001$; UR: $Z=-2.638$, $p=0.008$), they felt the least relaxed in CA compared with others (CR: $Z=-1.895$, $p=0.058$; UC: $Z=-2.781$, $p=0.005$; UR: $Z=-2.820$, $p=0.005$). They were also least confident in their time estimation for CA (CR: $Z=-2.539$, $p=0.011$; UC: $Z=-2.550$, $p=0.011$; UR: $Z=-3.297$, $p=0.001$). Our hypotheses **H3.1**, **H3.2** and **H4.1**, **H4.2** are all supported here.

We also found further evidence for **H3.1**, **H3.2** and **H4.1**, **H4.2**. Significant effects appeared on the TLX ratings of mental demand ($\chi^2=9.690$, $p=0.021$), performance ($\chi^2=12.627$, $p=0.005$) and effort ($\chi^2=15.426$, $p=0.001$), as shown in Figure 10: specifically, participants considered the mental demand in the CA condition to be much higher than

other three conditions (CR: $Z=-3.281$, $p=0.001$; UC: $Z=-2.774$, $p=0.006$; UR: $Z=-2.918$, $p=0.004$), the task performance to be poorer in CA than in others (CR: $Z=-3.171$, $p=0.002$; UC: $Z=-2.892$, $p=0.004$; UR: $Z=-2.990$, $p=0.003$), and the amount of effort to be higher in CA than others (CR: $Z=-3.348$, $p=0.001$; UC: $Z=-2.957$, $p=0.003$; UR: $Z=-2.926$, $p=0.003$).

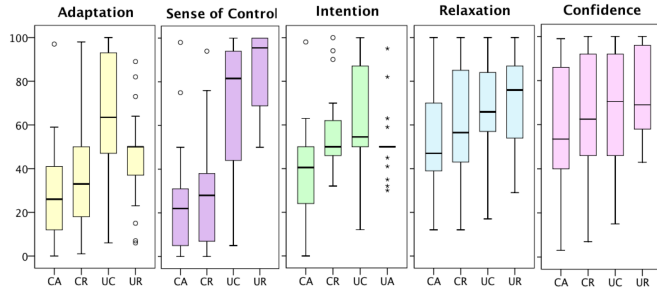


Figure 9: Ratings on 'Adaptation', 'Sense of Control', 'Intention', 'Relaxation' and 'Confidence' in Experiment 2

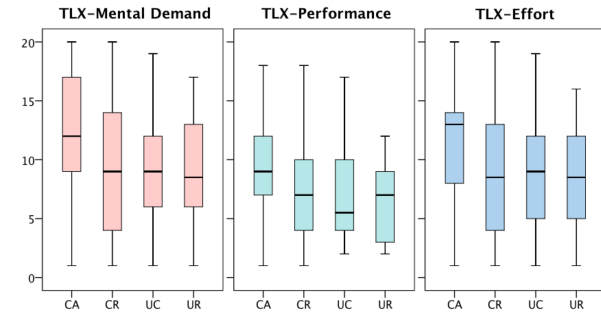


Figure 10: Ratings on TLX 'Mental Demand', 'Performance' and 'Effort' in Experiment 2

5.4.2. Outcome Binding

The analysis of outcome binding further demonstrates the effectiveness of our independent variable manipulation while supporting **H1.1** and **H1.2**. Again we used the non-parametric Friedman Test and the Wilcoxon Signed Ranks Test because the outcome binding data in the UR condition failed the Shapiro-Wilk Normality Test. Significant effect was found again ($\chi^2=46.893$, $p<0.001$), see Figure 11: outcome binding effect on the CA condition was significantly stronger than CR ($Z=-4.444$, $p<0.001$), UC ($Z=-4.067$, $p<0.001$) and UR ($Z=-6.262$, $p<0.001$). Both CR and UC conditions showed significantly stronger outcome binding effect than UR ($Z=-2.948$, $p=0.003$; $Z=-3.605$, $p<0.001$), while CR and UC had little difference statistically.

5.5. Discussion

In Experiment 2, as we predicted, the strongest outcome binding effect was observed in the CA

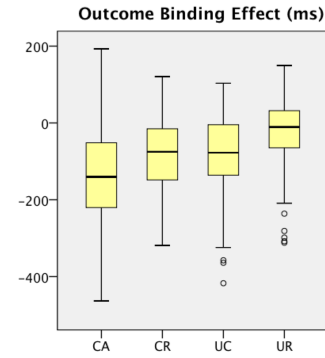


Figure 11: Outcome Binding Effect in Experiment 2

condition when the system presented arrhythmic auditory stimuli, while the binding effect was the mildest when participants controlled the pace. The binding effect in the CR condition was in between - significantly milder than in CA but stronger than in UR. This provides solid evidence that when users are not in control of the interaction pace, rhythmic intervals can preserve their sense of agency by providing a basis for temporal expectation.

Another interesting finding is that the binding effect appeared to be milder when there were 8 beeps in a round but became more salient with 7, 9 or 10 beeps. There might be an interaction between the number of beeps and the rhythm. We further analysed the outcome binding effect across conditions by grouping the rounds with 7, 8, 9 and 10 beeps separately, see results in Table 2. In the rounds with 8 beeps, significant binding effect was only found between UR and CA condition, but in other rounds, statistical difference was also seen between CR and CA, UC and CA, UR and UC, and marginally between UC and CR. We also noticed that there was no significant difference between UR and CR when we separately analysed the binding effects according to number of beeps, even if significance did appear between the average binding effect in all rounds of UR and CR.

Beeps	CR-CA	UC-CA	UR-CA	UR-UC	UC-CR	UR-CR
7	$Z=-3.43$ $p=0.001$	-	$Z=-3.78$ $p<0.001$	$Z=-3.43$ $p=0.001$	$Z=-1.96$ $p=0.050$	-
8	-	-	$Z=-2.55$ $p=0.011$	-	-	-
9	$Z=-2.10$ $p=0.037$	$Z=-2.58$ $p=0.010$	$Z=-3.23$ $p=0.001$	$Z=-2.16$ $p=0.031$	-	-
10	$Z=-2.65$ $p=0.008$	$Z=-3.41$ $p=0.001$	$Z=-2.84$ $p=0.004$	-	-	-

Table 2: Binding effects with different number of beeps

One possible explanation for this interaction effect is that when participants were listening to an uncertain

number of beeps, they automatically started to 'group' those signals to make it easier to attend to, and a group of four beats might be the most common pattern they had experienced. Since 8 beeps could be split into two 4-beep groups to fit a temporal expectation, this could mitigate the binding effect, indicating a preserved sense of agency. The predictability in the CR condition would allow participants to form temporal expectation, but because the number of beeps in each round was randomised, accumulated binding effect only emerged over time. Therefore, in mixed-initiative system design, if it is not possible to present the system's behaviours in a strictly rhythmic manner, we could consider grouping them with a regular temporal pattern in order to mitigate the reduced sense of control that results from the irregularity of single events.

6. FURTHER DISCUSSION

There are several limitations of this study and its findings. Firstly, these controlled tasks are a highly simplified form of mixed-initiative interaction. Most real systems have more complex behaviours and require more complex user decisions. Simply applying our findings in mixed-initiative system design may not be as effective as observed in these experiments. Secondly, the timescale of interaction intervals in these tasks ranged between 300ms - 2000ms, which is a relatively low granularity in human-computer interaction. There is not yet any evidence that our findings will be applicable on larger timescales. We are now doing further studies to investigate the two points above. The idea is to contextualise the findings from current research in a Programming-by-Example system. One possible scenario is to manipulate the timing of a series of decision making actions between users and an intelligent spreadsheet, which could (be perceived to) dynamically infer users' intention and update its formula. Another possible direction is to study if users' sense of agency would be altered differently when such an intelligent system asserts different levels of responsibility under a certain timing pattern. Both scenarios involve a back-and-forth initiative taking process on a greater time scale (1000ms - 5000ms). Thirdly, most participants in our experiments were not expert in mixed-initiative interaction, and they might have limited knowledge and expectation about such systems compared with experienced users or developers. We know that expectation plays a large role during initial allocation of responsibility [44; 45], which would influence how much effort users devote and how much control they assume. If this study were repeated with more expert participants, or if we introduced the study as testing an intelligent interface that is going to take over

control from time to time, it is likely that we would observe different effects.

7. CONCLUSION

When both users and system can take the initiative, time coordination of back-and-forth interaction becomes a key issue in system design. Users typically expect transition of control to happen just in time, without any noticeable overlap (where they try to reclaim the agency taken by the system) or gap (when neither assumes responsibility). Violating such expectations, whether received positively or as negative frustration, will trigger a process of reevaluation and redistribution of efforts and responsibility, potentially impairing the transition of control. To solve this problem, we have explored the effects of timing on users' perception of agency, hypothesising that rhythmic flow patterns during interaction can positively affect users' perceived agency, entrained behaviours, performance and relaxation, while arrhythmic patterns can be damaging on all these aspects. We designed and carried out two within-subjects experiments, one using visual stimuli and the other using auditory stimuli, that support our hypotheses. The major contributions of this study are: it establishes a research framework for HCI that draws on social psychology and neuropsychology; it demonstrates the importance of timing during mixed-initiative interaction; and it provides a quantitative measure of user sensitivity to the handover of initiative on a micro timescale. Our work suggests further research directions, to contextualise these findings within real applications, and to test whether they will generalise to a broader range of timescales. We hope that resulting insights, if used to inform mixed-initiative system design such as Programming-by-Example and end-user automation, will facilitate back-and-forth interaction with inference-based components of interactive systems.

8. ACKNOWLEDGEMENT

The authors would like to thank Professor Ian Cross (Centre for Music and Science, University of Cambridge), Dr David Coyle (School of Computer Science, University College Dublin) and Dr David Greatrex (Centre for Music and Science, University of Cambridge) for providing this research with expertise, advice and support.

REFERENCES

- [1] Cypher, A., (1991) Eager: Programming repetitive tasks by example. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. New Orleans, LA, USA, 27 April - 2 May 1991. New York, NY, USA: ACM, 33-39.

- [2] Darragh, J.J., Witten, I.H. & James, M.L. (1990) The reactive keyboard: A predictive typing aid. *Computer*, 23(11). 41-49.
- [3] Casner, S.M., Hutchins, E.L. & Norman, D. (2016) The challenges of partially automated driving. *Communications of the ACM*, 59(5). 70-77.
- [4] Horvitz, E. (1999) Principles of mixed-initiative user interfaces. In: *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*. Pittsburgh, PA, USA, 15-20 May. New York, NY, USA: ACM, 159-166.
- [5] Shneiderman, B. (1992) *Designing the user interface: strategies for effective human-computer interaction* (Vol. 3). MA: Addison-Wesley Reading.
- [6] Coyle, D., Moore, J., Kristensson, P.O., Fletcher, P., & Blackwell, A. (2012) I did that! Measuring users' experience of agency in their own actions. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. Austin, TX, USA, 5-10 May 2012. New York, NY, USA: ACM, 2025-2034.
- [7] Bratman, M. (1999) *Faces of intention: Selected essays on intention and agency*. Cambridge, UK: Cambridge University Press.
- [8] McCann, H. (1998) *The works of agency: On human action, will, and freedom*. Ithaca, NY, USA: Cornell University Press.
- [9] Haggard, P., & Tsakiris, M. (2009) The experience of agency feelings, judgments, and responsibility. *Current Directions in Psychological Science*, 18(4). 242-246.
- [10] Moore, J.W., & Haggard, P. (2008). Awareness of action: Inference and prediction. *Consciousness and Cognition*, 17(1). 136-144.
- [11] Limerick, H., Moore, J.W., & Coyle, D. (2015) Empirical evidence for a diminished sense of agency in speech interfaces. In: *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. Seoul, Republic of Korea, 18-23 April 2015. New York, NY, USA: ACM, 3967-3970.
- [12] Faratin, P., Sierra, C., & Jennings, N.R. (1998) Negotiation decision functions for autonomous agents. *Robotics and Autonomous Systems*, 24(3). 159-182.
- [13] Klingspor, V., Demiris, J., & Kaiser, M. (1997) Human-robot communication and machine learning. *Applied Artificial Intelligence*, 11(7). 719-746.
- [14] Harrison, C., Tan, D., & Morris, D. (2010) Skinput: Appropriating the body as an input surface. In: *Proceedings of the 2010 ACM SIGCHI Conference on Human Factors in Computing Systems*. Atlanta, GA, USA, 10-15 April 2010. New York, NY, USA: ACM, 453-462.
- [15] Iio, T., Shiomi, M., Shinozawa, K., Akimoto, T., Shimohara, K., & Hagita, N. (2011) Investigating entrainment of people's pointing gestures by robot's gestures using a WOz method. *International Journal of Social Robotics*, 3(4). 405-414.
- [16] Nash, C., & Blackwell, A. (2008) Realtime representation and gestural control of musical polytempi. In: *Proceedings of the 2008 Conference on New Interfaces for Musical Expression (NIME08)*. Genova, Italy: 5-7 June 2008. 28-33.
- [17] Chafe, C. (1993) Tactile audio feedback. In: *Proceedings of the International Computer Music Conference*. Tokyo, Japan, 10-15 September 1993. International Computer Music Association: 76-79.
- [18] Farrer, C., Bouchereau, M., Jeannerod, M., & Franck, N. (2008) Effect of distorted visual feedback on the sense of agency. *Behavioural Neurology*, 19(1-2). 53-57.
- [19] Ice, P., Curtis, R., Phillips, P., & Wells, J. (2007) Using asynchronous audio feedback to enhance teaching presence and students' sense of community. *Journal of Asynchronous Learning Networks*, 11(2). 3-25.
- [20] Stetson, C., Cui, X., Montague, P.R. and Eagleman, D.M. (2006) Motor-sensory recalibration leads to an illusory reversal of action and sensation. *Neuron*, 51(5). 651-659.
- [21] Moore, J.W., & Obhi, S.S. (2012) Intentional binding and the sense of agency: a review. *Consciousness and cognition*, 21(1). 546-561.
- [22] Desantis, A., Roussel, C., & Waszak, F. (2011) On the influence of causal beliefs on the feeling of agency. *Consciousness and Cognition*, 20(4). 1211-1220.
- [23] Fujioka, T., Trainor, L. J., Large, E.W., & Ross, B. (2009) Beta and gamma rhythms in human auditory cortex during musical beat processing. *Annals of the New York Academy of Sciences*, 1169(1). 89-92.
- [24] Rohenkohl, G., Cravo, A.M., Wyart, V., & Nobre, A.C. (2012) Temporal expectation improves the quality of sensory information. *The Journal of Neuroscience*, 32(24). 8424-8428.
- [25] Arnal, L.H., & Giraud, A.-L. (2012) Cortical oscillations and sensory predictions. *Trends in Cognitive Sciences*, 16(7). 390-398.
- [26] Summerfield, C., & Egner, T. (2009) Expectation (and attention) in visual cognition. *Trends in Cognitive Sciences*, 13(9). 403-409.
- [27] Wyart, V., Nobre, A.C., & Summerfield, C. (2012) Dissociable prior influences of signal probability and relevance on visual contrast sensitivity. *Proceedings of the National Academy of Sciences*, 109(9). 3593-3598.
- [28] Burgoon, J.K. (1995) Cross-cultural and intercultural applications of expectancy violations theory. *Intercultural Communication Theory*, 19. 194-214.
- [29] Burgoon, J.K. (1978) A communication model of personal space violations: Explication and an initial test. *Human Communication Research*, 4(2). 129-142.
- [30] Burgoon, J.K., White, C.A., & Greene, J.O. (1997) Researching nonverbal message production: A view from interaction adaptation theory. *Message Production: Advances in Communication Theory*. 279-312.

- [31] Patel, A.D. (2007) *Music, language, and the brain*. Oxford, UK: Oxford University Press.
- [32] Clayton, M., Sager, R., & Will, U. (2005) In time with the music: The concept of entrainment and its significance for ethnomusicology. *European meetings in ethnomusicology*, 11. 3-142.
- [33] Jones, M.R., & Boltz, M. (1989) Dynamic attending and responses to time. *Psychological Review*, 96(3). 459-491.
- [34] Schegloff, E.A. (1992) Repair after next turn: The last structurally provided defense of intersubjectivity in conversation. *American Journal of Sociology*, 97(5). 1295-1345.
- [35] Knight, S., Spiro, N. & Cross, I. (2017) Look, listen and learn: Exploring effects of passive entrainment on social judgements of observed others. *Psychology of Music*, 45(1). 99-115.
- [36] Hon, N., Poh, J.H., & Soon, C.S. (2013) Preoccupied minds feel less control: Sense of agency is modulated by cognitive load. *Consciousness and cognition*, 22(2). 556-561.
- [37] Libet, B., Gleason, C.A., Wright, E.W., & Pearl, D.K. (1983) Time of conscious intention to act in relation to onset of cerebral activity (readiness potential): The unconscious initiation of a freely voluntary act. *Brain*, 106. 623-642.
- [38] Haggard, P., Clark, S., & Kalogeras, J. (2002) Voluntary action and conscious awareness. *Nature Neuroscience*, 5(4). 382-385.
- [39] Gill, S.P. (2012) Rhythmic synchrony and mediated interaction: towards a framework of rhythm in embodied interaction. *AI & society*, 27(1). 111-127.
- [40] Nowicki, L., Prinz, W., Grosjean, M., Repp, B.H., & Keller, P.E. (2013) Mutual adaptive timing in interpersonal action coordination. *Psychomusicology: Music, Mind, and Brain*, 23(1). 6-20.
- [41] Boker, S.M., Rotondo, J.L., Xu, M., & King, K. (2002) Windowed cross-correlation and peak picking for the analysis of variability in the association between behavioral time series. *Psychological methods*, 7(3). 338-355.
- [42] Hawkins, S., Cross, I., & Ogden, R. (2013) Communicative interaction in spontaneous music and speech. *Music, Language and Interaction*. 285-329.
- [43] Hart, S.G., & Staveland, L. E. (1988) Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. *Advances in psychology*, 52. 139-183.
- [44] Fisek, M.H., Berger, J., & Norman, R.Z. (1995) Evaluations and the formation of expectations. *American Journal of Sociology*, 101(3). 721-746.
- [45] Fisek, M.H., Berger, J., & Norman, R.Z. (1991) Participation in heterogeneous and homogeneous groups: A theoretical integration. *American Journal of Sociology*, 97(1). 114-142.